

# cnSchema

## 开放中文知识图谱的schema

丁力 博士

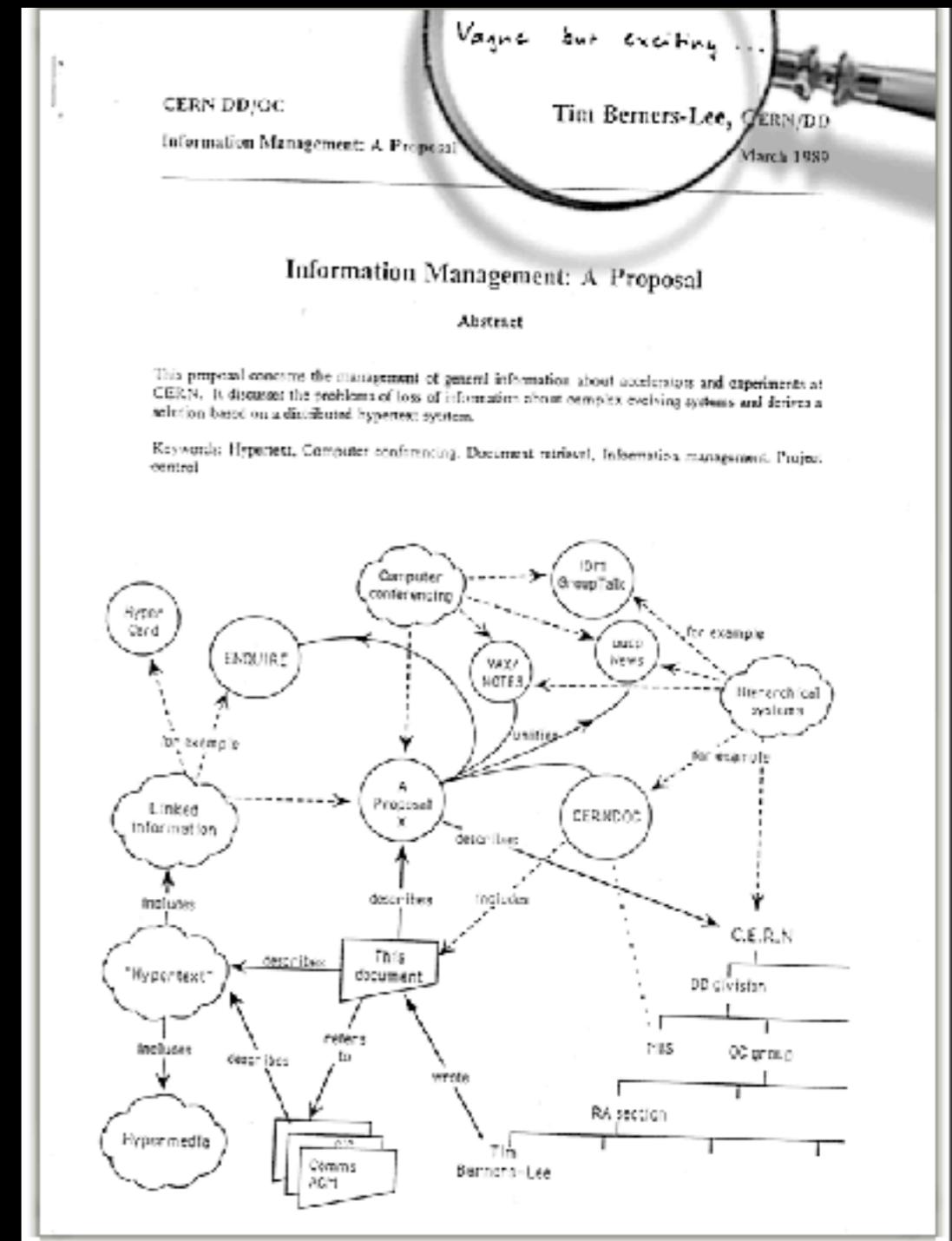
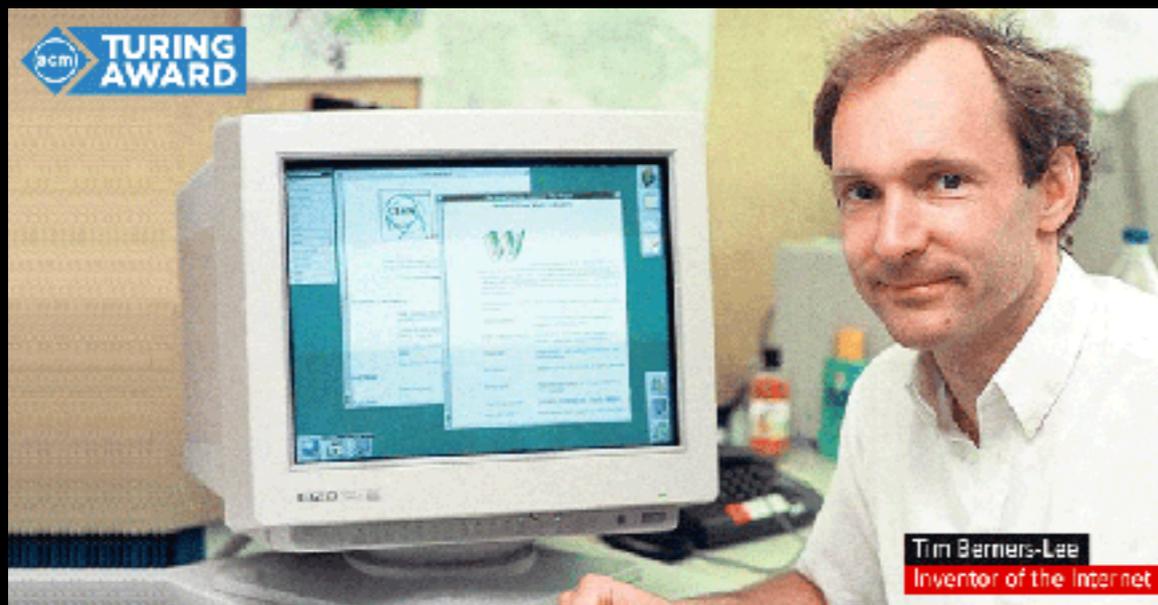
海知智能CTO OpenKG发起人

dl@ruyi.ai

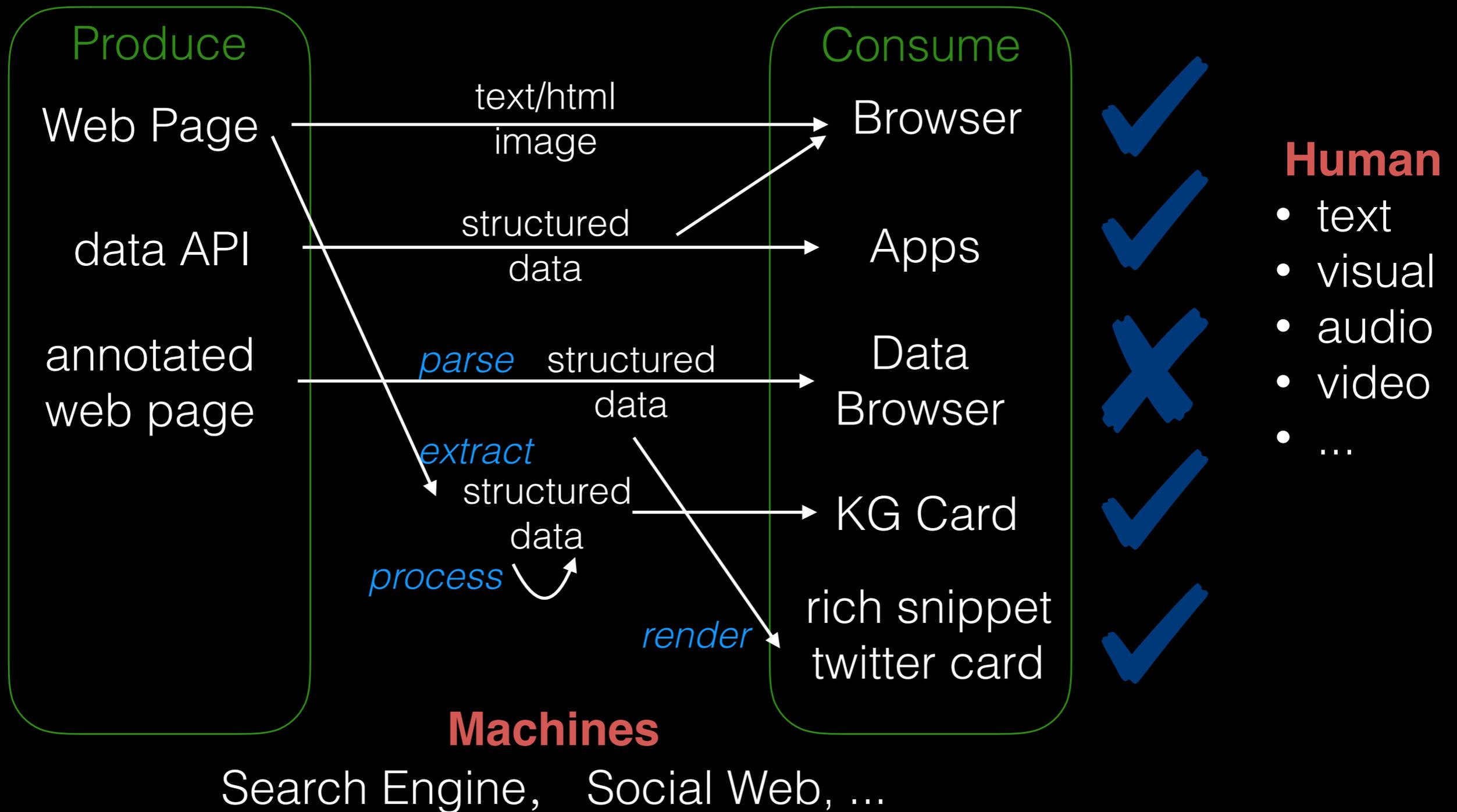
CCKS2017, 成都, 2017-08-29

# Tim Berners-Lee's vision for the Web (1989, 1998)

The Web was designed as an information space, with the goal that it should be useful not only for **human**-human communication, but also that **machines** would be able to participate and help.



# Web for Human, powered by machines



# Schema Matters

- scope of schema

- Data Model
- Syntax
- Vocabulary
- Identifiers for Object

- the killer app

- standard first (since 1996) ? MCF, RDF, OWL, FOAF, RSS, ...
- data first (since 2004) ? Swoogle, DBpedia, Linked Data, Open Government Data
- **consumer first** (since 2007) ! searchMonkey, [schema.org](http://schema.org)

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  分类 "@type": "MusicRecording",
  "@id": "http://musicbrainz.org/recording/3566e45",
  "name": "Back in the U.S.S.R.",
  "producer": {
    "@type": "Person",
    "name": "George Martin"
  },
  属性 "duration": "PT2M43S",
  "recordingOf": {
    "@type": "MusicComposition",
    "name": "Back in the U.S.S.R.",
    "iswcCode": "T-010.140.236-1"
  }
}
</script>
```

格式

ID

名称

属性值

关系

**schema.org**

<http://schema.org/MusicRecording>

# R.V. Guha's Schema.org promotes great adoption (2011)

.... A longstanding goal of the semantic web initiative is to get webmasters to make the **structured data directly available on the web** .... Learning from these earlier attempts has guided the development of [schema.org](http://schema.org)

- Clear incentives, SEO
- One vocabulary understood by major search engines
- make it easy for webmasters

**Nikon D3200 24.2 MP Digital SLR Camera with 18-55mm ... - Target**  
[www.target.com/.../nikon-d3200...digital-slr-camera-wi...](http://www.target.com/.../nikon-d3200...digital-slr-camera-wi...) - Cached  
 ★★★★★ Rating: 5 - 3 reviews - \$599.99  
 Nikon D3200 24.2MP Digital SLR Camera with 18-55mm VR Lens Black .... save \$150 on a Nikon 55-300mm AF-S DX ED VR NIKKOR Zoom lens (online item# ...

**Nikon D3200 Digital SLR Camera 18-55mm G VR Zoom Lens 24.2 ...**  
[www.ebay.com/Cameras & Photo > Digital Cameras](http://www.ebay.com/Cameras%20%26%20Photo/Digital%20Cameras) - Cached  
 ★★★★★ Rating: 5 - 48 votes - \$489.95  
 Nikon D3200 Digital SLR Camera + 18-55mm G VR Zoom Lens 24.2 MP Black USA in Cameras & Photo, Digital Cameras | eBay.

**Rich Snippets**

FIGURE 4A: MAJOR SITES THAT HAVE PUBLISHED SCHEMA.ORG

CATEGORY	SITES
News	nytimes.com, guardian.com, bbc.co.uk
Movies	imdb.com, rottentomatoes.com, movies.com
Jobs / Careers	careerjet.com, monster.com, indeed.com
People	linkedin.com, pinterest.com, familysearch.org, archives.com
Products	ebay.com, alibaba.com, sears.com, cafepress.com,
Video	youtube.com, dailymotion.com, frequency.com, vin
Medical	cvs.com, drugs.com
Local	yelp.com, allmenus.com, urbanspoon.com
Events	whereevent.com, meetup.com, zillow.com, eventful
Music	last.fm, myspace.com, soundcloud.com



12+ Million Web Sites

A de facto vocabulary for structured data on the web

Found On 30% Pages

**schema.org**

# How Schema.org Works

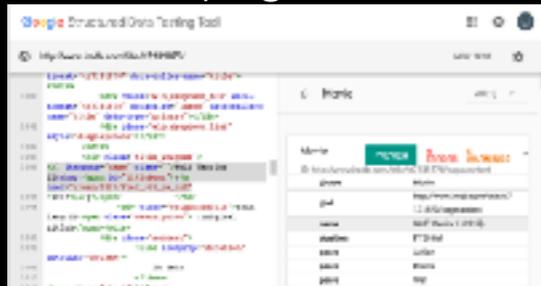
1. publish web page



2. check Movie schema



3. enrich page with annotation



4. consume rich snippet



A large screenshot of the IMDb page for 'Wolf Warrior II (2017)'. The page features a search bar at the top, navigation tabs, and a prominent movie title with a 6.8 rating. Below the title is a movie poster and a trailer video player. The trailer is titled '有人吗? Anyone home?' and has a duration of 1:51. Below the trailer is a 'Get Showtimes &amp; Tickets' button and a brief description of the movie: 'China's deadliest special forces operative settles into a quiet life on the sea. When sadistic mercenaries begin targeting nearby civilians, he must leave his newfound peace behind and...'. The page also includes a 'Watchlist' button and a 'Rate This' button.

# How Schema.org Works

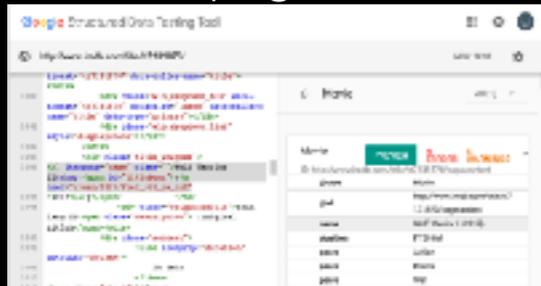
1. publish web page



2. check Movie schema



3. enrich page with annotation



4. consume rich snippet

A large screenshot of the schema.org website. The top navigation bar is dark red with the 'schema.org' logo, a search bar, and links for 'Home', 'Schemas', and 'Documentation'. The main content area is white and features the 'Movie' schema details. It includes the canonical URL 'http://schema.org/Movie', the hierarchy 'Thing > CreativeWork > Movie', a description 'A movie.', and usage information 'Between 10,000 and 50,000 domains'. A table lists properties from the 'Movie' schema.

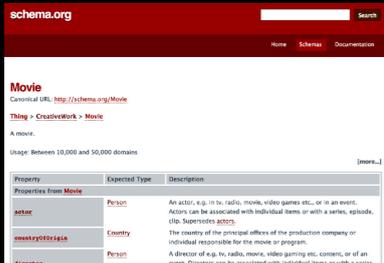
Property	Expected Type	Description
<b>Properties from <u>Movie</u></b>		
<u>actor</u>	<u>Person</u>	An actor, e.g. in tv, radio, movie, video games etc., or in an event. Actors can be associated with individual items or with a series, episode, clip. Supersedes <u>actors</u> .
<u>countryOfOrigin</u>	<u>Country</u>	The country of the principal offices of the production company or individual responsible for the movie or program.
<u>director</u>	<u>Person</u>	A director of e.g. tv, radio, movie, video gaming etc. content, or of an event. Directors can be associated with individual items or with a series,

# How Schema.org Works

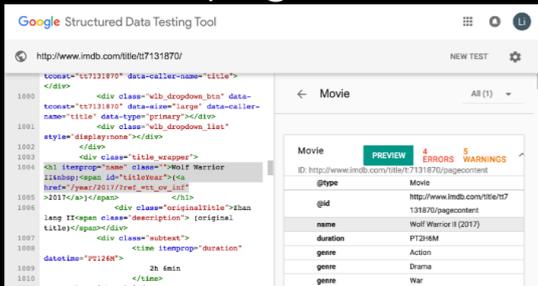
## 1. publish web page



## 2. check Movie schema



## 3. enrich page with annotation



## 4. consume rich snippet



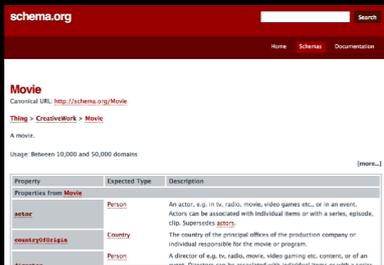
A large screenshot of the Google Structured Data Testing Tool interface. The top bar shows the URL 'http://www.imdb.com/title/tt7131870/'. The left pane displays HTML code with line numbers 1000-1015, showing the 'Movie' schema annotations. The right pane shows a preview of the rich snippet for the movie 'Wolf Warrior II (2017)'. The preview includes a 'PREVIEW' button, '4 ERRORS' and '5 WARNINGS' indicators, and a table of properties: @type (Movie), @id (http://www.imdb.com/title/tt7131870/pagecontent), name (Wolf Warrior II (2017)), duration (PT2H6M), genre (Action, Drama, War), and datePublished (2017-07-27). An 'image' property is also listed with a URL to an Amazon image.

# How Schema.org Works

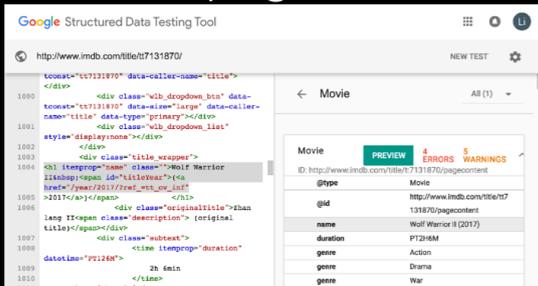
## 1. publish web page



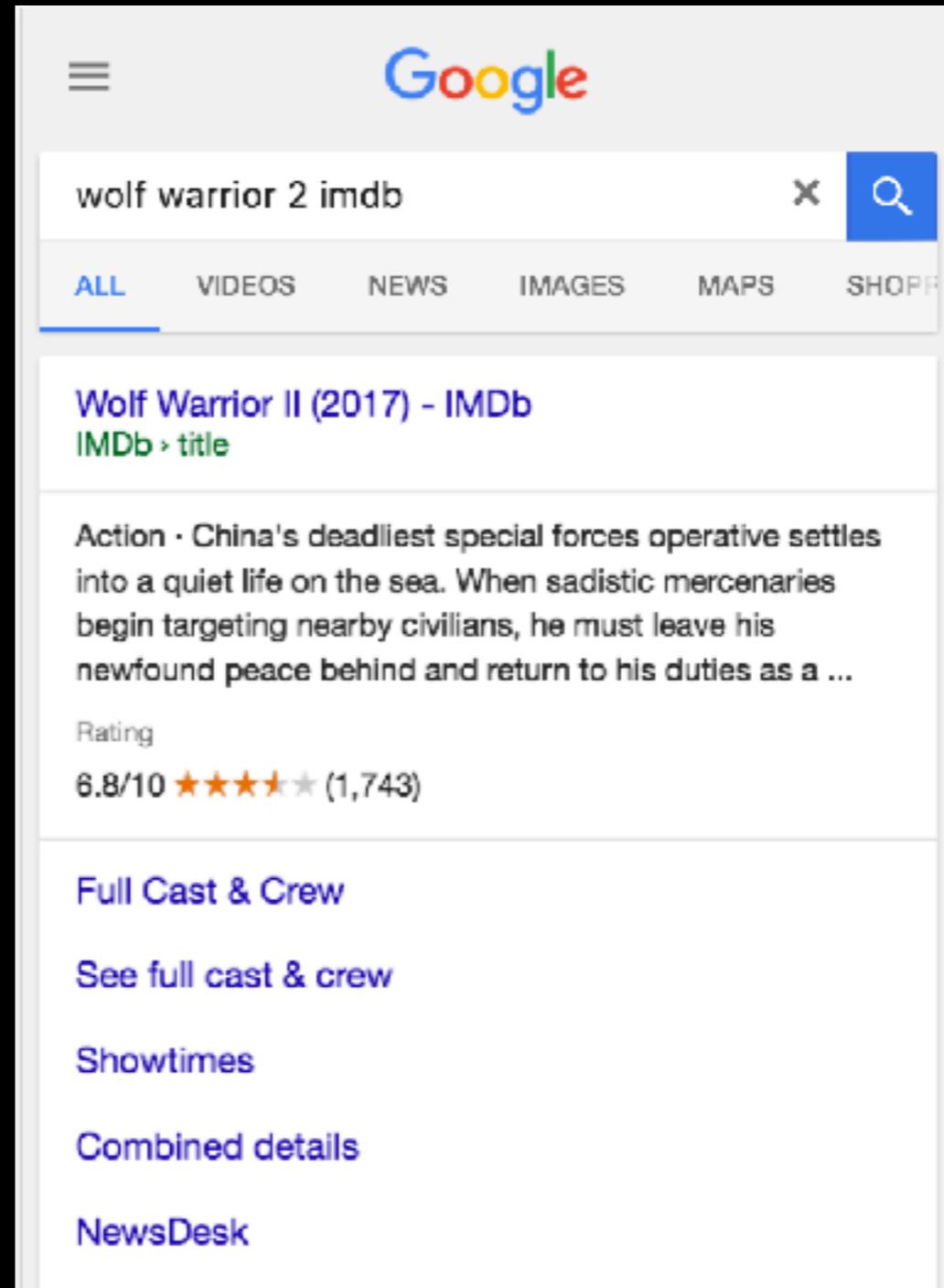
## 2. check Movie schema



## 3. enrich page with annotation



## 4. consume rich snippet





Money is useless unless you spent it

and

Schema is useless unless use it with data

# Schema for Bot and KG

## 产品经理

“用户关心什么问题？”  
“我的产品需要哪些数据？”

## 开发者

“接口字段怎么理解？”  
“KG存储结构和查询怎么写？”

## 数据发布者 / 采集者

“我有一些数据？”  
“我的数据哪里需要改进？”  
“让人找到并用到我的数据？”

学习问答意图模版

苏东坡的父亲是谁？

@kg.person 的 @cns.father 是谁？



查属性名定义

描述数据源结构



# cnSchema.org

cnSchema

项目

文档

词汇表

关于

OpenKG.CN



## 欢迎访问cnSchema.org

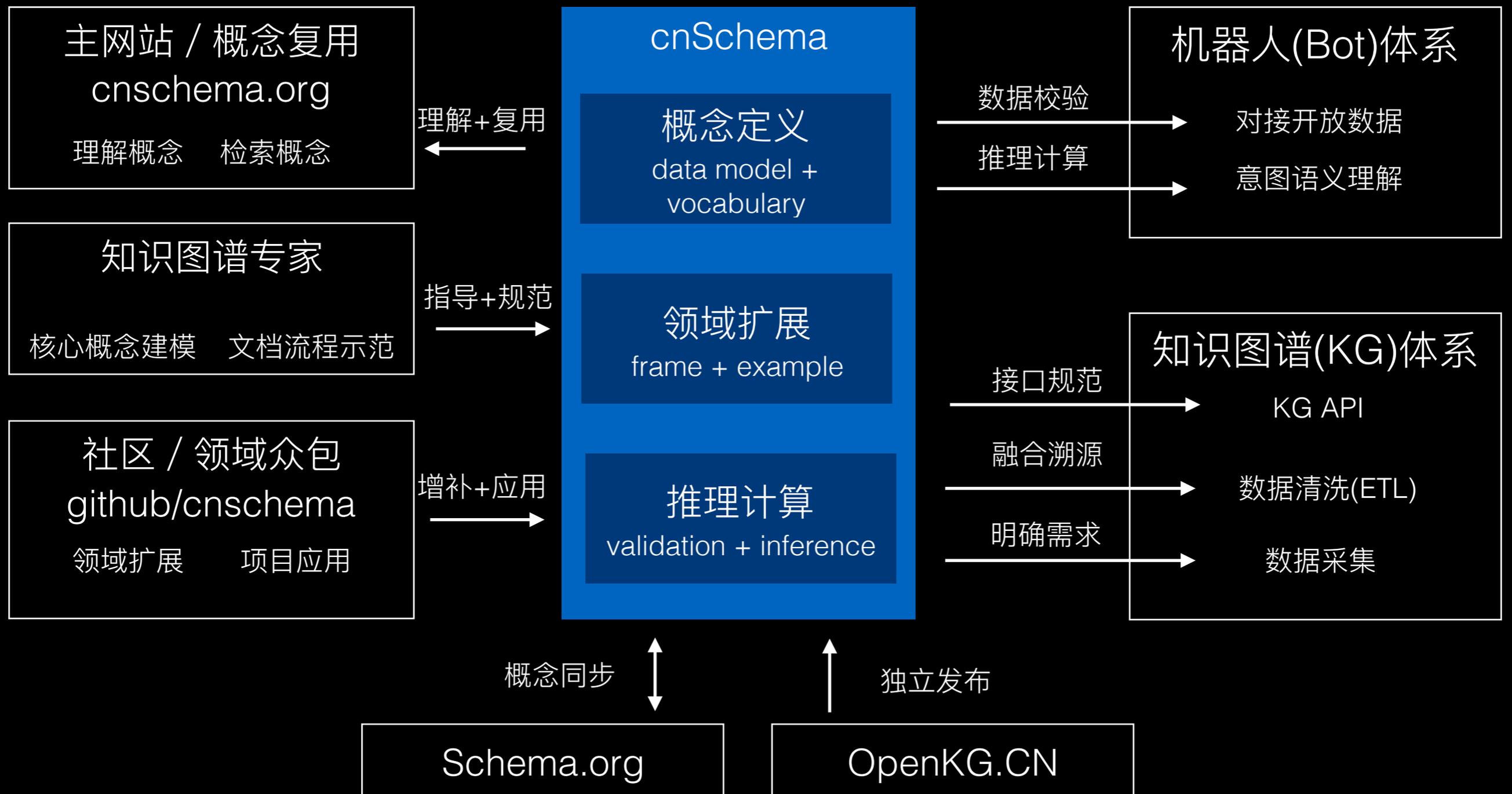
[cnSchema.org](#)是一个基于社区维护的开放的知识图谱Schema标准。cnSchema的词汇集包括了上千种概念分类(classes)、数据类型(data types)、属性(propertities)和关系(relations)等常用概念定义, 以支持知识图谱数据的通用性、复用性和流动性。结合中文的特点, 我们复用、连接并扩展了[Schema.org](#), [Wikidata](#), [Wikipedia](#)等已有的知识图谱Schema标准, 为中文领域的开放知识图谱、聊天机器人、搜索引擎优化等提供可供参考和扩展的数据描述和接口定义标准。通过cnSchema, 开发者也可以快速对接上百万基于[Schema.org](#)定义的网站, 以及Bot的知识图谱数据API。

[开始使用](#)

# cnSchema

- 源自[schema.org](https://schema.org)，由OpenKG自主发布
- 基于中文，支持全球中文市场
- 面向Bots应用
- 开放的schema
- 由知识图谱专家指导

# cnSchema 生态体系



# 案例分析：佛学人物问答

(东南大学 + ruyi.ai)

# 微信公众号对话界面



Bot 引擎

意图理解

任务完成

回复生成

# 知识图谱问答插件

(ruyi.ai)



知识图谱问.. kg.person 老师、师傅、导师-studentOf

数据统计

粉丝管理

对话场景

- kg.thing
- kg.person

词典实体

导入知识库

机器人设置

技能插件

素材管理

老师、师傅、导师-studentOf

化名-pseudonym

学生、弟子、徒弟、门徒-student

法号、法名-dharmaName

成就、主要成就-accomplishment

作品、代表作-authorOf

籍贯-ancestralHome

死亡日期-deathDate

宗派-schoolsOfBuddhism

朝代、所在年代、时代-dynasty

职业、身份-occupation

\* 用户说

- @kg.person:entity 的师傅是谁
- @kg.person:entity 的老师是谁
- @kg.person:entity 的导师是谁
- @kg.person:entity 是谁教的
- @kg.person:entity 是谁的学生

查看更多

机器人答

微信 硬件

微信回复

# 佛学知识图谱数据示例

(东南大学)

Bot 引擎

意图理解

任务完成

回复生成

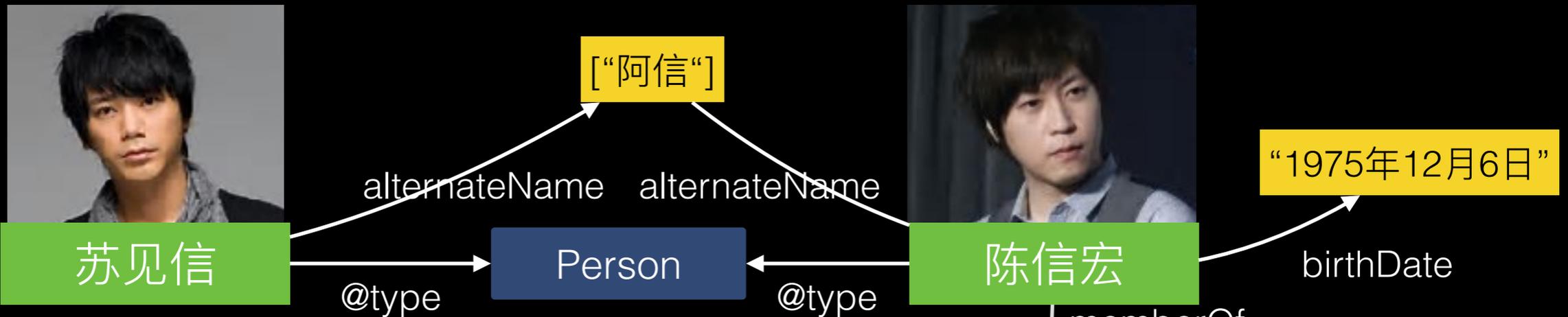
```
- alternateName: [
  "弘一法师; 晚晴老人; 演音; 李息霜; 李岸",
  "原名李叔同, 幼名成蹊、广侯, 名息, 学名文涛, 字叔同、息霜, 号漱筒、演音等, 别署甚多。",
  "中国书法家、文学家和著名佛教僧侣",
  "弘一法师"
],
+ image: [...],
schoolsOfBuddhism: "佛教",
ancestralHome: "浙江平湖",
DBpediaLink: "http://zh.dbpedia.org/resource/李叔同",
+ relatedPage: [...],
name: "李叔同",
- dharmaName: [
  "演音",
  "漱筒、弘一、晚晴"
],
abstractBaidu: "李叔同, 又名李息霜、李岸、李良, 谱名文涛, 幼名成蹊, 学名广侯, 字息霜, 别号漱筒; 祖籍浙江平湖, 生于天津。中国话剧的开拓者之一, 在音乐、书法、绘画和戏剧方面, 都颇有造诣。从日本留学归国后, 担任过教师、编辑之职, 后剃度为僧, 法名演音, 号弘一, 晚号晚晴老人。",
- alias: [
  "弘一大师",
  "弘一法师"
],
- student: [
  "丰子恺",
  "刘质平",
  "李鸿梁",
  "丰子恺"
],
+ birthName: [...],
@id: "<http://www.kg-buddhism.com/entity/李叔同>",
```

# cnSchema

## 典型应用场景

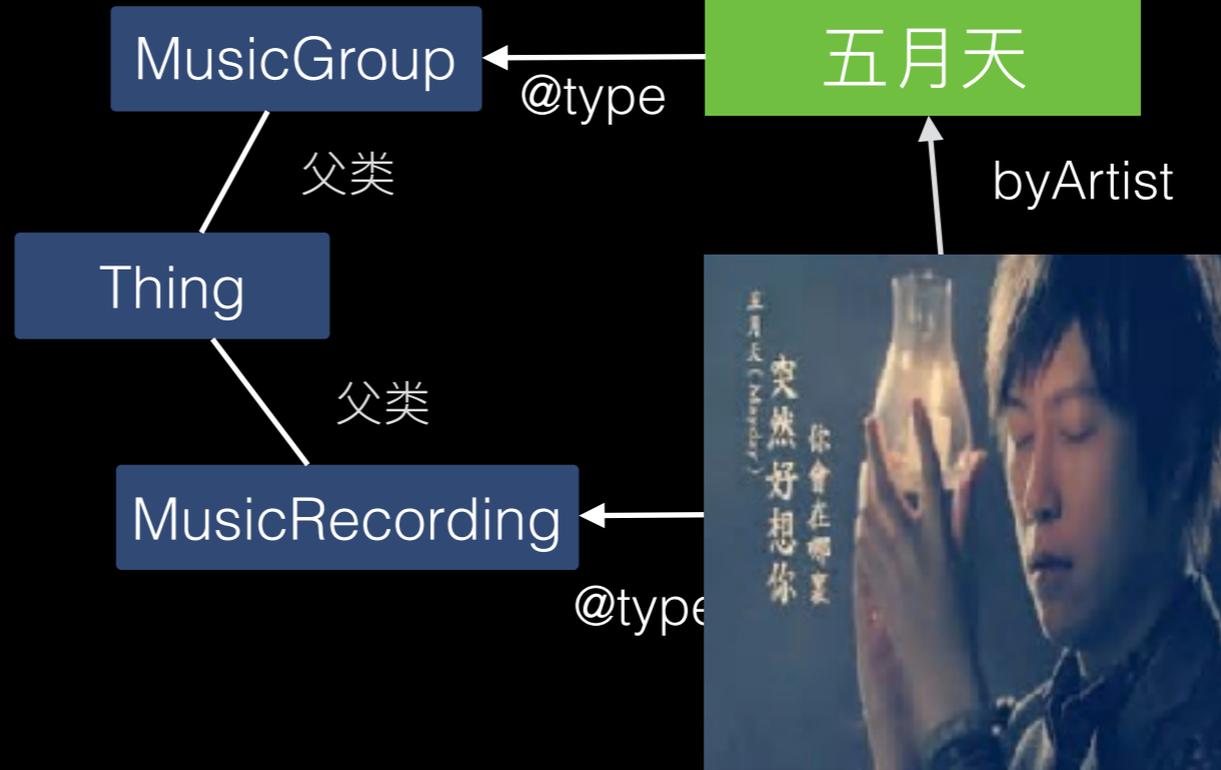
# Schema for Bots

## Entity Linking, query, and task completion



Entity Linking using web statistics

P(苏见信) 92,300 RESULTS  
 P(陈信宏) 53,600 RESULTS  
 P(陈信宏,阿信) 25,800 RESULTS  
 P(苏见信,阿信) 16,200 RESULTS



五月天的热门歌曲 TOP100		
全部播放		
01	突然好想你	78521104
02	如果我们不曾相遇	75900993
03	后来的我们	70560900
04	我不愿让你一个人	57430920
05	好好 (想把你写成一首歌)	
06	动画电影《你的名字。》中文主题曲 / Song About You	56201085

- 我要听“突然好想你”
- 我要听“阿信”的歌
- 放一个“突然好想你你会在哪里过得快乐或委屈”
- “阿信”是“五月天”里最老的吗?



“...突然好想你 你会在哪里，过得快乐或委屈...”

# Schema for Web Data Extraction and Fusion



采集+清洗



```
{ "statedIn": "虾米",  
  "playCount": 16948661,  
  ... }
```

usage statistics

- \* 信息来源、用户播放数等数据不能融合
- \* 融合结果溯源，支持数据可信度分析

融合



```
{ "mergedFrom": [  
  { "statedIn": "虾米",  
    "playCount": 16948661,  
    ... },  
  { "statedIn": "网易",  
    "playCount": 2731000,  
    ... } ],  
  ...  
}
```

provenance



采集+清洗

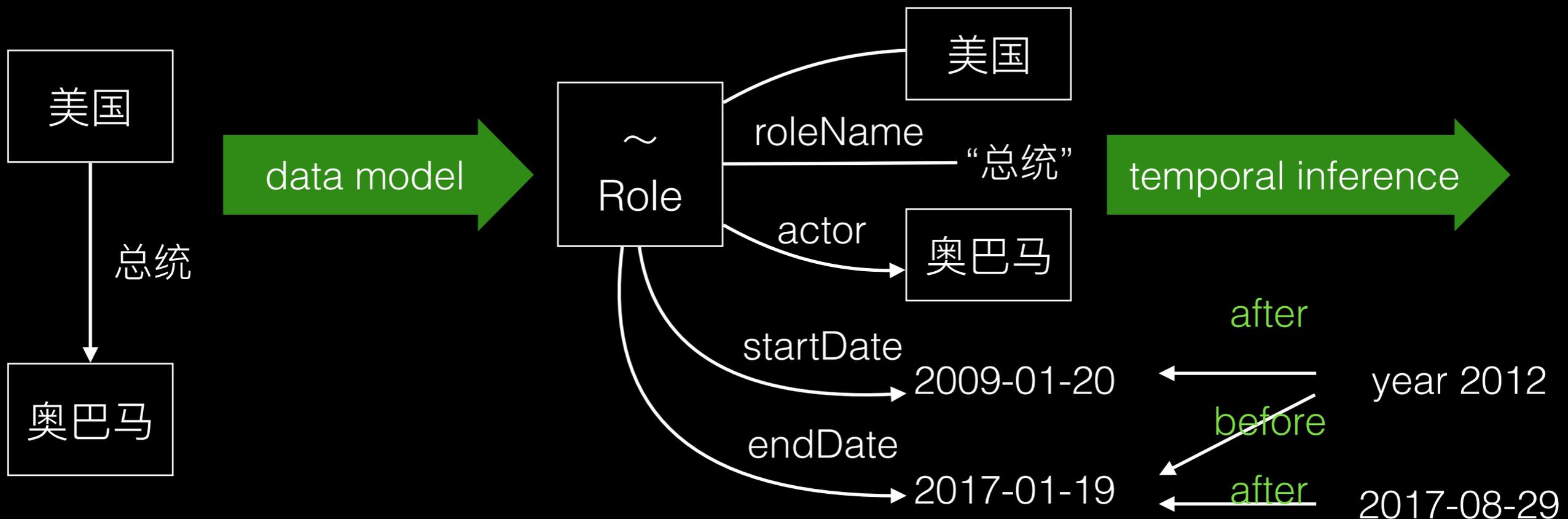


```
{ "statedIn": "网易",  
  "playCount": 2731000,  
  ... }
```

“十年在虾米的评价如何？”

# Schema for KG Data Modeling and Inference

- 现在的美国总统是谁?
- 2012年美国总统是谁?



# Schema for Extending Chinese Concepts

- 拿“籍贯”这个属性来说，只有中国人有，所以schema.org没有收录。wikidata有收录，<https://www.wikidata.org/wiki/Property:P66>，因此使用 `ancestralHome`
- 有些中文属性更难翻译到贴切的英文概念，只能采用拼音。

C	D	E	F	G	H
规范属性名	中文属性名	schema.org属性名	wikidata属性名		cnschema属性名
name	姓名	<a href="#">name</a>			
alternateName	别名	<a href="#">alternateName</a>			
description	简介	<a href="#">description</a>			
image	图片	<a href="#">image</a>			
keywords	标签	<a href="#">keywords</a>			
birthPlace	出生地	<a href="#">birthPlace</a>	<a href="#">P19</a>		
birthDate	出生日期	<a href="#">birthDate</a>	<a href="#">P569</a>		
deathDate	死亡日期	<a href="#">deathDate</a>	<a href="#">P570</a>		
deathPlace	死亡地	<a href="#">deathPlace</a>	<a href="#">P20</a>		
placeOfBurial	墓地、安葬地		<a href="#">P119</a>		placeOfBurial
homeLocation	家庭地址	<a href="#">homeLocation</a>			
alumniOf	毕业院校	<a href="#">alumniOf</a>	<a href="#">P69</a>		
ancestralHome	籍贯		<a href="#">P66</a>		ancestralHome
occupation	职业、身份		<a href="#">P106</a>		occupation
cnProfessionalTitle	职称				cnProfessionalTitle
fieldOfWork	领域、专业		<a href="#">P101</a>		fieldOfWork
academicMajor	高校专业		<a href="#">P812</a>		academicMajor
ethnicGroup	民族、民族族群		<a href="#">P172</a>		ethnicGroup
nobleFamily	家族		<a href="https://www.wikidata.org/wiki/Property:P119">https://www.wikidata.org/wiki/Property:P119</a>		nobleFamily
religion	宗教信仰、信仰		<a href="#">P140</a>		religion
memberOfPoliticalParty	政党、党派		<a href="#">P102</a>		memberOfPoliticalParty
courtesyName	字、表字		<a href="#">P1782</a>		courtesyName
artName	号、自号、别号、又号		<a href="#">P1787</a>		artName
templeName	庙号		<a href="#">P1785</a>		templeName
posthumousName	谥号、私谥		<a href="#">P1786</a>		posthumousName
pseudonym	化名		<a href="#">P742</a>		pseudonym
birthName	原名		<a href="#">P1477</a>		birthName
familyName	姓氏	<a href="#">familyName</a>	<a href="#">P734</a>		familyName

cnSchema

任务与进展汇报

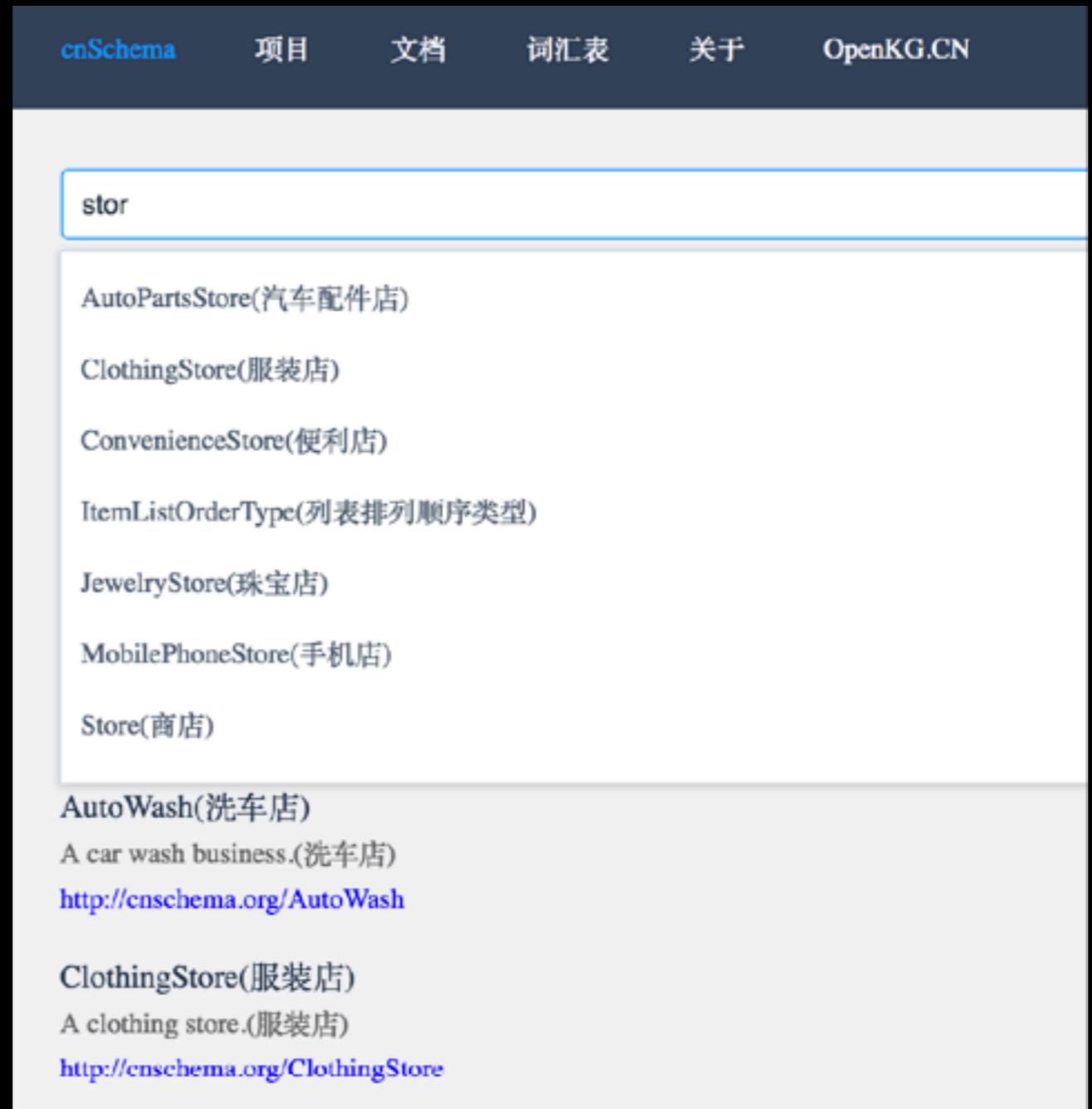
# TASK 1: 核心概念的中文翻译与链接

- 本项工作基于 [schema.org](http://schema.org) 核心词汇，由清华大学，浙江大学，复旦大学，东南大学，海知智能，以及社区志愿者共同完成
- TODO: 优化翻译和链接质量，增补中文场景的数据样例

```
{
  "@id": "http://cnschema.org/birthPlace",
  "alternateName": [
    "出生地"
  ],
  "category": "property",
  "description": "The place where the person was born.",
  "descriptionZh": "此人出生的地方。",
  "name": "birthPlace",
  "nameZh": "出生地点",
  "schemaorgUrl": "http://schema.org/birthPlace",
  "supersededBy": "",
  "wikidataName": "place of birth",
  "wikidataUrl": "http://www.wikidata.org/entity/P19",
  "wikipediaUrl": "https://en.wikipedia.org/wiki/Birth_place"
},
```

# TASK2: 概念搜索与语义分析

- 概念的全文检索以及自动完成API. Swoogle's ontology dictionary is back
- TODO: 基于分布式表示的概念相似度计算, 支持概念语义搜索, 例如“超市”也能搜到“商店”
- TODO: 基于实际使用的概念排序, 促进热门概念的复用

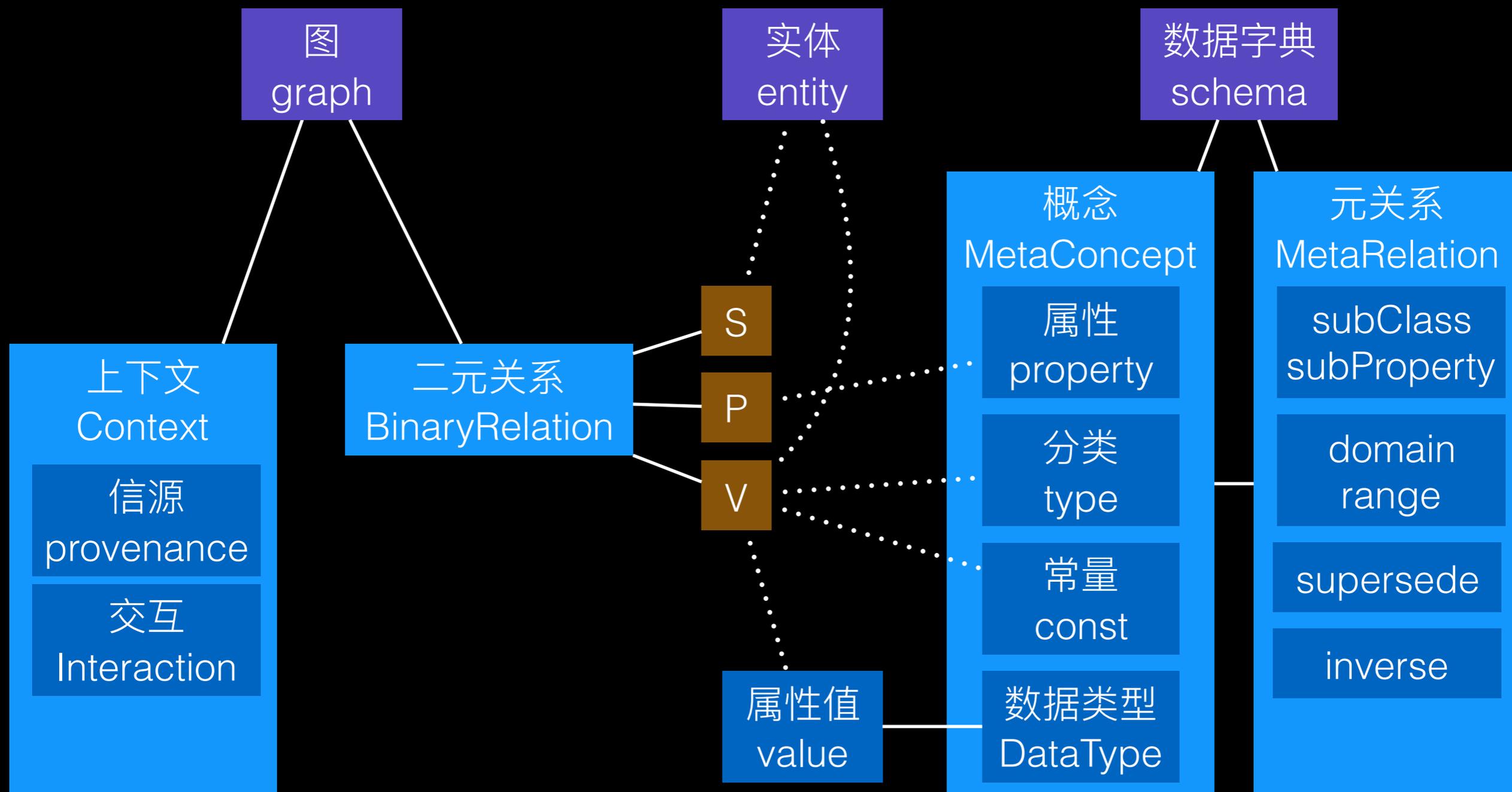


The screenshot shows the cnSchema website interface. At the top, there is a navigation bar with links for 'cnSchema', '项目', '文档', '词汇表', '关于', and 'OpenKG.CN'. Below the navigation bar is a search input field containing the text 'stor'. The search results are displayed in a list format, showing various ontology classes with their English and Chinese names and their URIs. The results are as follows:

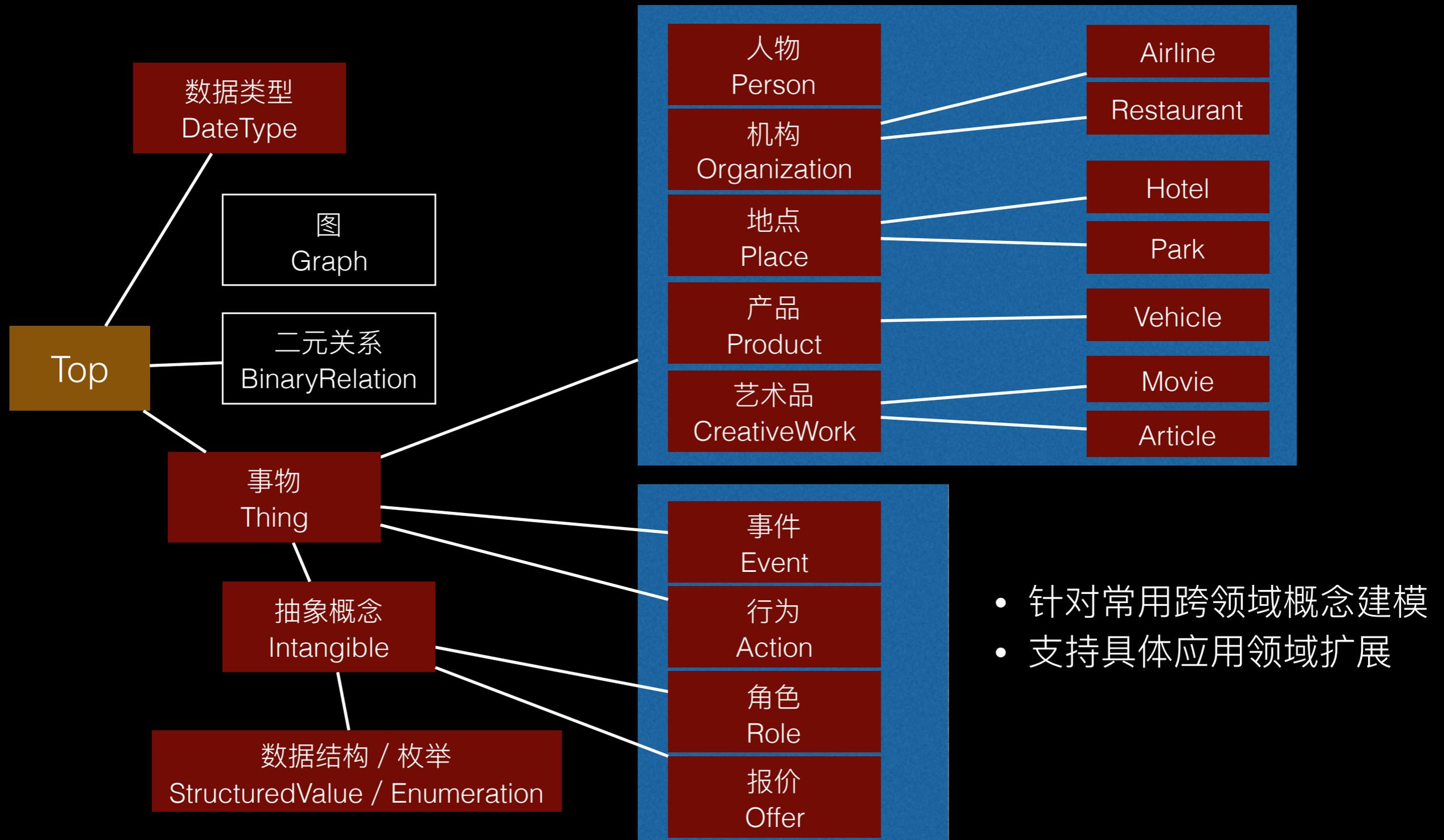
- AutoPartsStore(汽车配件店)
- ClothingStore(服装店)
- ConvenienceStore(便利店)
- ItemListOrderType(列表排列顺序类型)
- JewelryStore(珠宝店)
- MobilePhoneStore(手机店)
- Store(商店)
- AutoWash(洗车店)  
A car wash business.(洗车店)  
<http://cnschema.org/AutoWash>
- ClothingStore(服装店)  
A clothing store.(服装店)  
<http://cnschema.org/ClothingStore>



# Task4: KG基本数据模型的需求分析与结构优化



# Task5: 核心schema的简化



# Task6: 知识图谱API设计

The screenshot shows the documentation for the 'EntityCoreAPI - Lookup Entities' endpoint. On the left is a sidebar with navigation links: 'EntityCoreAPI', 'Get Entity', 'Lookup Entities' (highlighted), 'EntityIndexAPI', 'Cypher Graph Query', 'ElasticSearch Query', 'EntitySyncAPI', 'Batch Update', 'Delete an entity', 'List Entities', and 'Update/Create an entity'. The main content area has a title 'EntityCoreAPI - Lookup Entities' and a description: 'redis lookup service。按名字 获取一组实体简略信息 (@id, @type, name, entityScore)'. Below this is a 'POST' method indicator and the endpoint path 'entities/'. A 'Parameter' table follows, with columns 'Field', 'Type', and 'Description'. The table contains one row for the 'names' parameter, which is a 'String[]' and is marked as '必须' (required) with the description 'list of Entity name, match whole word.'. Below the table is a 'Lookup Query' input field. At the bottom, a code block shows a JSON example: { "names": [ "华仔", "唱歌" ] }.

EntityCoreAPI - Lookup Entities

redis lookup service。按名字 获取一组实体简略信息 (@id, @type, name, entityScore)。

POST

entities/

Parameter

Field	Type	Description
names	String[]	<b>必须</b> list of Entity name, match whole word.

Lookup Query

```
{
  "names": [
    "华仔",
    "唱歌"
  ]
}
```

<https://lidingpku.github.io/kgapi/apidoc> (征求意见)

# Task 7: KG领域应用、Schema扩展、 以及知识图谱应用示范流程

- 佛学人物 (东南大学+)

- 音乐 (海知智能+)

- 影视 (清华+)

- 新闻 (清华+)

- 家谱 (上图)

- 金融 (文因互联+)

- 电商 (浙江大学+)

- 企业组织 (复旦大学+)

- 菜谱 (豆果+)

- 中医药 (中医药研究所+)

- 地点 (天津大学+)

- 一带一路 (海知智能 + 若干大学和科研机构)

- ...



# 感谢cnSchema志愿者

- cnSchema是OpenKG正在努力的一个方向
- cnSchema由来自清华大学、浙江大学、北京大学、复旦大学、东南大学、南京大学、英国阿伯丁大学等十多所国内外高校的计算机科学专家，以及微软亚洲研究院、海知智能、狗尾草科技、文因互联等企业所共同发起、建立并维护。
- cnSchema得到Schema.Org的负责人R.V. Guha和Dan Brickley，以及语义网创始人之一Jim Hendler教授的指导和支持。

[cnschema@openkg.cn](mailto:cnschema@openkg.cn)

丁力 上海海知智能 CTO/博士  
陈华钧 浙江大学 教授  
漆桂林 东南大学 教授  
王昊奋 深圳狗尾草 CTO/博士  
谢殿侠 上海海知智能 CEO  
李涓子 清华大学 教授  
闫峻 微软亚洲研究院 研究员  
肖仰华 复旦大学 副教授  
鲍捷 文因互联 CEO/博士  
曾毅 中科院自动化所 研究员  
Jeff Pan 英国阿伯丁大学 教授  
邹磊 于彤  
张鹏 徐波  
侯磊 胡伟  
吴天星 徐常亮  
张大卫 徐艺  
仲亮靓 王梁  
张宇轩 邓淑敏  
王宇 杨平京  
孙娜 董孙宏璐  
陈旭 郭唯  
Kevin Xin



加入cnschema志愿者